DATA SHEET

# Data Science in Practice

## HOW TO BECOME A PIVOTAL CERTIFIED DATA SCIENTIST

### DELIVERY METHODS

- Instructor-led

### COURSE DURATION

- Five days of instructor-led classroom training
- 50% lecture and small exercises, 10% demos and walk-throughs, 40% extended Data Science project exercises

### TARGET AUDIENCE

- Experienced data analysts and data engineers willing to work hard to achieve superior Pivotal Data Science skills.
- Anyone else who wants to learn about data science using the Pivotal product stack.

### PREREQUISITES

- Willingness to participate in a demanding, high-intensity training experience
- Comfort with data analytic technologies a plus (Statistics, mathematics, machine learning, SQL, R, Python)
- Have a basic understanding of virtualization and massive parallel processing concepts

### PRICING

Please visit our website at gopivotal.com/training

### MORE INFORMATION

On-site training is also available for customers who prefer to bring a Pivotal Certified Instructor to their own facilities For additional information about on-site classes, including facility requirements, contact education@gopivotal.com

## COURSE OVERVIEW

This course is designed to give the student hands-on experience with the Pivotal products related to performing Pivotal Data Science projects. Given the diverse and varying nature of customer implementations this course will focus on the main aspects of a Data Science project within Pivotal: Pivotal Greenplum DB, pSQL, MADlib, GPText, PivotalHD, HAWQ, PivotalR, pyMADlib with extra units covering Alpine Chorus and Visualization. Students are introduced to the need for big, fast data and its role in modern business applications; The course will provide hands on experience using Pivotal Greenplum DB, pSQL, MADlib, GPText, Apache Hadoop, Pivotal HD, HAWQ, Alpine Chorus, PivotalR, PL/R, pyMADlib, PL/Python, and several visualization tools such as Gephi, D3, and Tableau. This course will introduce and use, but does not include extensive training on, pSQL, R, Python. Further, this course will provide attendees with an opportunity to explore several intense Data Science projects that have been converted into extensive Data Science exercises. This course does not teach Installation, Configuration, and Management of any of the products.

## COURSE OBJECTIVES

After attending this course, students should be able to:

- Summarize the distinguishing characteristics of each Pivotal product and tool, and be able to describe the most beneficial aspects from a Data Science perspective;
- Evaluate and demonstrate hands-on practical skills with each product and tool;
- Investigate, assess, and apply their knowledge to practical data science problems;
- Practice Data Science problem solving techniques to their respective endeavors.

As a result of attending the course, the Data Scientist will be able to confidently utilize the Pivotal product set and related technologies to analyze large data sets.

# COURSE MODULES

## 1. INTRODUCTION

## 2. DATA SCIENCE OVERVIEW

- Data Science: The Big Picture
- Driving Forces
- What Does a Data Scientist Do
- The Process of Data Science
- What Does Pivotal bring to the Story

## 3. PIVOTAL OVERVIEW

- Pivotal Corporate Overview
- The Pivotal Big Data Suite
  1. Pivotal Greenplum DB
  2. Pivotal GPText
  3. MADlib
  4. Pivotal HD
  5. Pivotal Data Dispatch (PDD)
  6. Pivotal on Virtualized Hardware
  7. Pivotal HAWQ
  8. Pivotal eXtension Framework (PXF)
  9. Pivotal Analytics Workbench
  10. Pivotal GemFire
  11. Pivotal GemFireXD
  12. Spring by Pivotal
  13. Spring XD
  14. Pivotal Labs and Pivotal Data Labs

## 4. PIVOTAL GREENPLUM DB REVIEW INCLUDING INLINE LABS

- Essentials
- Getting Started and Inline Lab Exercise
- Intro to pSQL and Inline Lab Exercises
  1. Creating Tables
  2. Distributions and Partitioning
  3. Indexes
  4. External Tables and Loading Data
- Unloading Data
- Analyze
- Explain and Analyze
- Vacuum
- Monitoring

## 5. ADVANCED SQL

- Explore and Inline Lab Exercise
- Joins and Inline Lab Exercise
- Arrays and Array Aggregates and Inline Lab Exercise
- Window Functions and Inline Lab Exercise
- Other Functions and Inline Lab Exercise
- User Defined Functions (UDF's)
- User Defined Aggregates (UDA's)
- Data Science Exercise

## 6. MADLIB INCLUDING INLINE LABS

- MADlib Basics
- Advanced MADlib
- Data Science Exercise

## 7. GPTEXT INCLUDING INLINE LABS

- Introduction to GPText
- Working with GPText
  1. Creating Indicies
  2. Searching a GPText Index
  3. GPText Analyzers
- Text Analytics with GPText
- NLP: Practical Examples
- NLP: Practical Examples with NLTK
- Putting it all together
- Data Science Exercise

## 8. APACHE HADOOP AND THE HADOOP ECOSYSTEM INCLUDING INLINE LABS

- Apache Hadoop Overview
  1. Core Component: HDFS
  2. Core Component: MapReduce
  3. Map Reduce: Writing a Job
- Hadoop Ecosystem
  1. Hadoop Streaming
  2. Pig

# COURSE MODULES cont.

## 9. PIVOTAL HD AND HAWQ INCLUDING INLINE LABS

- Intro to Pivotal HD and HAWQ
- Getting Started with HAWQ
- Working with HAWQ
- External Tables: file, gpfdist, web
- External Tables: PXF
- Loading and Unloading Data and Inline Lab Exercises
  1. Loading and Unloading using Copy
  2. Loading and Unloading using Insert
  3. Loading and Unloading using gpfdist / gpload / external tables
- Data Science Exercise

## 10. GEMFIRE AND GEMFIRE XD (OPTIONAL)

## 11. ALPINE CHORUS

- Chorus: Collaborative Big Data Analytics
- Chorus Live Walk-through

## 12. R AND PYTHON

- PivotalR
- PL/R
- pyMADlib
- PL/Python
- Data Science Exercise

## 13.VISUALIZATION

- Tableau
- R
- Python
- Exercises

## 14. HAWQ TEXT ANALYTICS EXERCISE AIRLINE PRICE OPTIMIZATION EXERCISE GENE SEQUENCING EXERCISE

- HAWQ Text Analytics Exercise
- Airline Price Optimization Exercise
- Gene Sequencing Exercise

Pivotal™

At Pivotal our mission is to enable customers to build a new class of applications, leveraging big and fast data, and do all of this with the power of cloud independence. Uniting selected technology, people and programs from EMC and VMware, the following products and services are now part of Pivotal: Greenplum, Cloud Foundry, Spring, GemFire and other products from the VMware vFabric Suite, Cetas and Pivotal Labs.

Pivotal 1900 S Norfolk Street San Mateo CA 94403 **goPivotal.com**

GoPivotal, Pivotal, and the Pivotal logo are registered trademarks or trademarks of GoPivotal, Inc. in the United States and other countries. All other trademarks used herein are the property of their respective owners.
© Copyright 2013 Go Pivotal, Inc. All rights reserved. Published in the USA. PVTL-DS-114-04/13